The role of data modelling in a modern geological survey

M. NAYEMBIL[1*] and G. BAKER[2]

[1]Data Science, Informatics – British Geological Survey, United Kingdom, mln@bgs.ac.uk
[2] Data Science, Informatics - British Geological Survey,
*presenting author

Abstract

Data models are key elements in understanding the meaning (semantics) of data and communicating the information requirements for geoscientific-environmental research. The British Geological Survey (BGS) models its data to provide an understanding of: the organisational information requirements and its communication; the nature of its data independent of their physical representations and to facilitate the access of the data across multiple outlets. Generally, the components of an optimal model are that it has: structural validity, simplicity, expressibility, nonredundancy, shareability, integrity and can be presented diagrammatically. However, trade-offs are sometimes necessary to avoid the loss of simplicity in trying to achieve greater expressibility in a data model. The paper explores the role of data modelling at the British Geological Survey (BGS) in developing an integrated geoscientific data model, a component of a multi-tiered data architecture.

1 Introduction

As a geological survey organisation (GSO), the BGS maintains a huge breadth of data and information as part of it's role as a National Geoscience Data Centre (NDGC) and also for it's research. It holds the physical and digital data from geoscience investigations undertaken by the itself and also data donated by external bodies. The vast majority of this data has a spatial component and attributed with metadata and database constraints for the indexed data. The digital data is held in relational databases, open and/or native file formats.

Data management plays a critical role at the BGS in enabling, the organising of it's data for; integration, discovery, access, download and use. As an organisation, the BGS has used professional standards, methodologies and industry standard software packages to capture, store, integrate, validate and make accessible its data in differing ways through many platforms for both an internal and external audience. The data types range from a highly variable legacy to newly acquired geological data to include: mapping, geochemistry, geophysics, geotechnics, site investigation, outputs from 3D modelling, sensor networks (new data streams) including data donated from external bodies (mandated or by agreement).

To enable the understanding, capture, storage, integration and use of such diverse and large data types and their long-term re-use, the BGS uses data modelling as a key component in it's Geoscience data hub (GDH). Over many years and still evolving, BGS has developed an integrated multi-tiered data architecture for its geoscientific data, at the heart of which is an integrated geoscientific data model (GOM) that is extendable, encapsulating the different and many data elements with their business rules.

This paper explores data modelling at the BGS, it's role in developing an integrated geoscientific data model and in defining an overall data architecture that supports the wide range of users and applications: computation applications, data delivery applications, web sites, web services, smartphone apps, temporal web applications whilst also remaining useful and useable for future activities.

2 Data Modelling

At the core of building an integrated geoscientific data model, is understanding the different data elements, standardising them, how they relate to one another and their representation. This process of defining and analysing data requirements needed to support business processes referred to as Data Modelling takes you through the design stages: conceptual, logical and physical represenations of the model.

There are many scholarly texts that define a data model, but the overidding and simplistic thought is that a data model represents reality, whichever reality it is that you are modelling. The data model organises data elements and standardises how the data elements relate to one another and communicates this information in a diagramatic representation to interested parties (e.g. system developers). Also it is the evidence to share for a similar but separate representaion to facilitate easy sharing of the modelled facts, enabling the interoperability of those facts. In the context of the BGS, these data elements relate to boreholes, borehole interpretations, geochemcial analysis, site investigations, geophysical sampling, geological field observations, photographs, borehole core, seismic interpretaions, 3D models amongst many others.

According to (Fleming and Von Halle, 1989) an optimal data model should have the following: structural validity, simplicity, expressibility, nonredundancy, shareability, integrity and can be presented diagrammatically. The main aim of data models is to support the development of information systems by providing the definition and format of data. According to (West and Fowler, 1999) "if this is done consistently across systems then compatibility of data can be achieved. If the same data structures, semantics/classifications are used to store and access data then different applications can share data.

The task of creating systems and interfaces which involves, to build, operate and maintain the systems can often be expensive, so it's important to get it right, with flexibility and re-usable. It's therefore imperative to have high quality, extendable and shareable data models, so that they don't become bottlenecks for the organisation but rather support it. As a result, the BGS undertakes data modelling to define new entities and extending existing ones in its integrated geoscientific data model. We undertake conceptual design: identifying the important entities, relationships and attributes in our data without physical considerations; Logical Design; then translate these important entities, attributes and relationships into a specific data model (e.g. relational model) without other physical considerations and then Physical design: translate all of the above into a physical implementation of choice (e.g. Oracle RDBMS). The model is integrated to maximise the interrelationships of the datasets covering the various subject areas, through business rules, standards, common vocabularies and good design practices. For our data modelling, we use an industry standard tool "Embarcadero ER/Studio Data Architect" to design, document and share our data models. Figure 1 shows the data model of the core Borehole index. Because of this integrated data model, it enables us to discover data across different data types that are related through common attributes (e.g. location, measured properties, lithostratigraphy and/or lithological units). It provides a level of consistency across the data structures, with a singular meaning of an attribute within a particular scope. Vocabulary terms are consistent across different parts of the model to provide a common meaning and understanding of geoscientific terms.

Vocabularies are a key component in developing a geoscientific model more so an integrated one. At the BGS, our vocabularies are, in effect, controlled vocabularies, they control the terms we wish to use in describing, and supporting the description of, scientific and other observations. They enable us to re-use and extract real value from our diverse geoscience-environmental datasets. For example; the ability to relate geology to groundwater, geophysics or engineering properties, gives us greater data mining and analytics possibilities to understand trends in the data to help answer scientific questions. They also allow for the understanding between differing scientists and clients and enable the data to be compared or integrated during scientific processes and re-used.

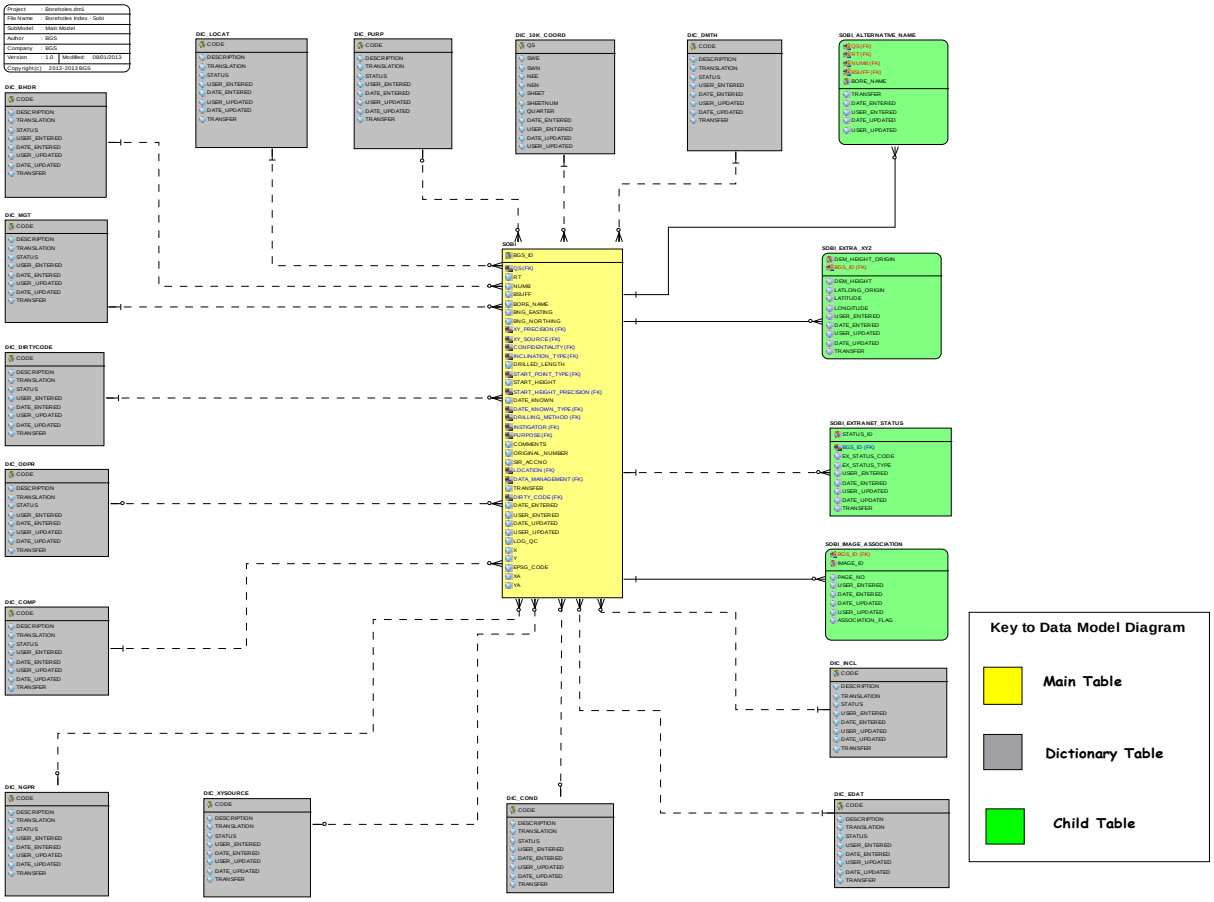

Figure 1: Data Model of the core Borehole Index (contains 1.3 million borehole records)

3 Data Architecture

Over many years, and still evolving BGS has developed a data architecture that is multi-tiered to hold and serve the varied data assets of the organisation. This data architecture referred to as the "Geoscience Data Hub" has several component parts to include: relational database systems both proprietary and open source (Oracle, MS SQL Server, MySQL, PostgreSQL), NoSQL databases – MongoDB and a file store on a corporate Storage Area Network – SAN that holds open or propriety file formats as either the original data formats and/or generated file formats for use in systems from other data stores. The data hub has as part of it, an integrated geoscientific data model (outlined in Section 2 above) for a majority of BGS's digital data that are appropriate to database management technology. It encapsulates the business of the organisation and its data interrelationships, with strong emphasis on the integration of spatial and non-spatial datasets.

The HUB has linkages to our file store/Storage Area Network for those datasets stored externally to the database, but with their primary metadata held within the HUB. There are also linkages to external systems that hold specialised datasets better managed by such standalone systems. It's a totality of an integrated environmental data HUB with components parts. The major business rules within the RDBMS component are captured, implemented, validated and fully documented. As a result of this implementation of the data architecture, it has become a definitive source of digital datasets in the BGS that are quality assured and have appropriate safeguards in place to minimise legislative compliance risk; e.g. data protection, freedom of Information issues and EIR. It also uniquely positions the BGS for our research, knowledge transfer, collaboration, products and service delivery challenges now and in the future. Figure 2 illustrates a high level data architecture with its

multi-tiered and hybrid component parts, whereas Figure 3 outlines the specific implementation of the data architecture in a relational database (RDBMS), the databases component in Figure 2.

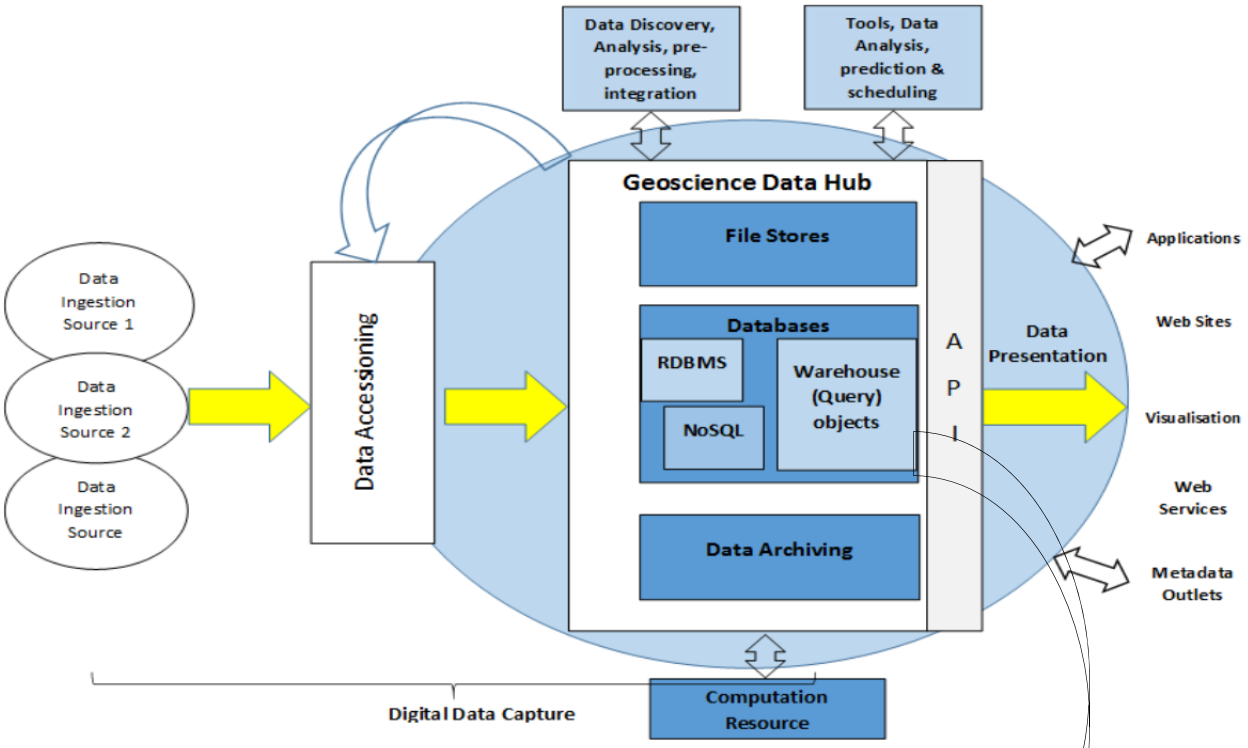

Figure 2: BGS High level multi-tiered data architecture and dataflow

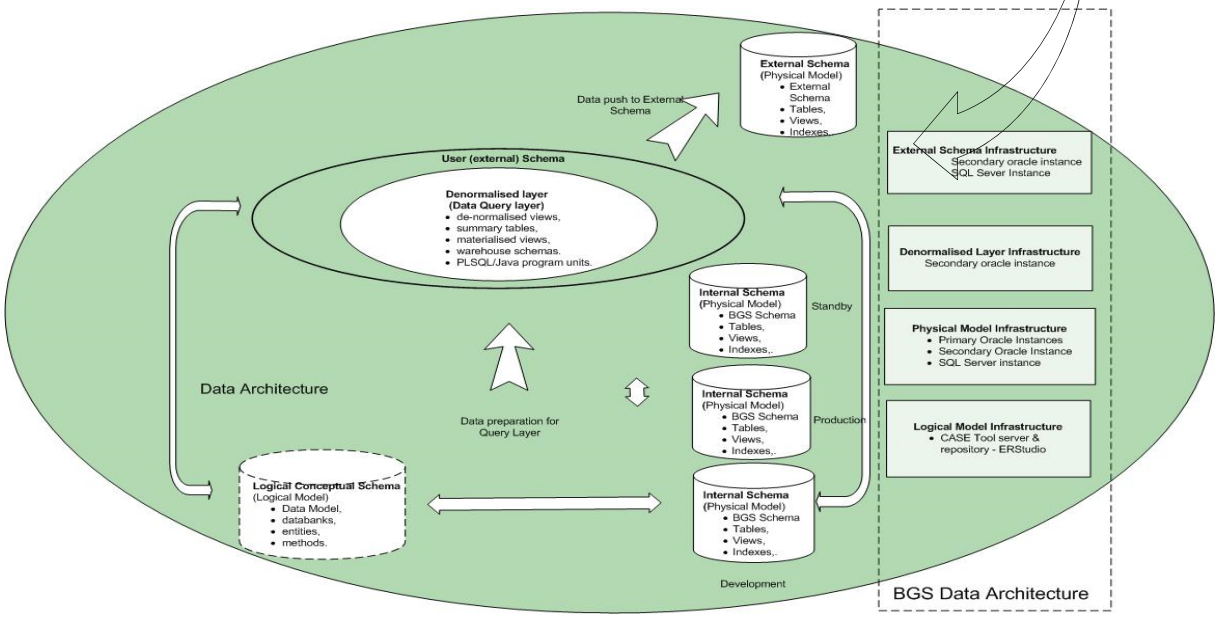

Figure 3: BGS Data Architecture specific to a RDBMS Implementation

The multi-tiered and hybrid data architecture in Figure 2, demonstrates the great variety in BGS's data types, but also the increasing volume and different formats of data it increasingly has to deal with as an organisation. We are continuously extending the data architecture to accommodate new data types especially from sensors networks, as we extend our data coverage to 4D time-series data from near-real time monitoring of earth processes to promote better scientific understanding of these and to assist with decision making on environmental impacts and in building sustainable global and smart cities.

At the core of this data resource is also metadata; we maintain a high level accession system as a level one catalogue of our data and information assets. We have a core implementation of ISO 19115 compliant metadata system with extensions and mappings to other international standards (e.g. GEMINI, INSPIRE) to enable us register our datasets with as many metadata gateways to allow for the easy discovery, download and use of the datasets. The RDBMS implementation contains about a hundred (100) separate corporate databases covering key geoscientific subject areas to include; boreholes, borehole logs, borehole core, geophysics, hydrogeology, landslides, geochemistry, geotechnics, marine, controlled vocabularies – e.g. Rock Classification Scheme, Lithostratigraphic Lexicon, Images, Palaeontology, mineral occurrence and statistics, 3D geological objects and sensor measurements.

Some of the benefits gained with our RDBMS implementation for our geoscientific data include:

- Maintain the integrity of the databases and data held within them in a centralised environment for our science and beyond.
- Maximise the interrelationships of the databases for an integrated corporate database covering various subject areas, through business rules, standards, common dictionaries and good design practices.
- Ensure easy access to the core objects, but also denormalized data for our many applications, products and services.
- Provide access to data from a single source
- Reduce lost knowledge or updates
- Reduce duplication
- Reduce organisation risk (Legislation/Legal Compliance)

3.1 Standards and Best Practice

As part of the architecture as outlined in Figure 3, we promote the use of standards/best practice for our systems development especially in data modelling and in the implementation of databases. These assist BGS in achieving consistency in the representation and meaning of its data models and databases, compatibility internally and externally, standardisation of data types, making it possible to share and communicate its data models and generally to promote interoperable systems internally and externally.

BGS have defined design standards for databasing, adapting international standards like ISO where applicable for in-house use: Some of these best practice cover areas to include: database design and documentation, design methodologies, formulation of data definitions, database objects naming conventions, vocabulary design standard, storage of geographic coordinates in a RDBMS.

As already indicated above, we maintain a huge catalogue of controlled vocabularies across our geoscience disciplines, mostly defined in-house by our domain experts with others adopted from international standards bodies (e.g. ISO, BS).

3.2 Denormalized “Query” Layer

To support the multi-tier information architecture at the BGS and to facilitate making the data held in our databases easily accessible to a host of applications both internally and externally (computation applications, delivery applications, web sites, web services, smartphone apps, temporal web applications), we have implemented a denormalized “Query” layer within our data architecture, akin to data warehousing techniques onto our databases (see Figure 3).

The query layer was built to tackle issues that have been afflicting BGS IT development for some time and are a common issue with integrated database systems with a great deal of interrelationships without a middle-tier (layer) implementation. Having applications built directly on correctly normalized database tables to enforce integrity, design methodology and only optimised for storage without considering how these tables may be used in different applications can lead to problems to include; poor query performance, complex data transformation embeded in the application layer for data presentation or complex application specific SQL queries to format the data as per the end-user requirement, the core database tables are not abstracted from the presenation layer,

inability to work on the core database objects if applications are directly linked, hence any subsequent developments on the core tables can affect the performance and robustness of directly linked applications, affecting the quality of the service provided to end-users. None of these issues are desirable in any system development, and so BGS implemented the 'Query Layer' concept as a series of denormalised objects (to divorce the complex SQL embeded in applications and to pre-create summary objects in the database, presenting these summary objects to applications). The database will handle the optimisation and specific objects can be created for specific applications and/or as export formats for other systems.

This has proved very successful and provides an abstraction layer of the data from our databases for our numerous external applications whilst at the same time supporting the internal science community and 3D modelling. We have also extended the query layer concept into the successful development of PropBase, a "data warehouse" system that extracts, transforms and loads data into a simplified data model from across BGS's heterogeneous property databases into a single data structure, so that the data is compatible and accessible from a single interface for internal delivery. This is then made accessible through a discovery and download tool including the provision of a web-service, enabling other software systems to directly interrogate the datasets to visualise them. We are also applying a similar methodolgy to time series data - 4D from sensors that we currently hold within our databases.

4 Application and Use

The implementation of a multi-tiered data architecture as described above with it's component parts, databases, filestores, query layer objects attributed with metadata at dataset and entity level has uniquely positioned the BGS for it's reserach, in undertaking knowledge transfer with other GSO's, collaboration, the development of data products and services, in building, storing and delivering 3D models, exposing it's data through web-services and different types of data visualisation. It has provided BGS the ability to re-use it's datasets across different products due to the standardisation of it's data structures and formats, and has led to a reduction in unnecessary data duplication with a clear data feed-in and out loop.

It has also provided the platform to now publish, share, review, redesign and extend the implementation of our existing designs for different platforms with our Open Geoscience Data Models inititiative (http://www.bgs.ac.uk/services/dataModels/home.html) at minimal expense and not overly compromise existing systems. The examples below show outputs generated from the data held in our Geoscience Data Hub as part of a multi-tiered data architecture.

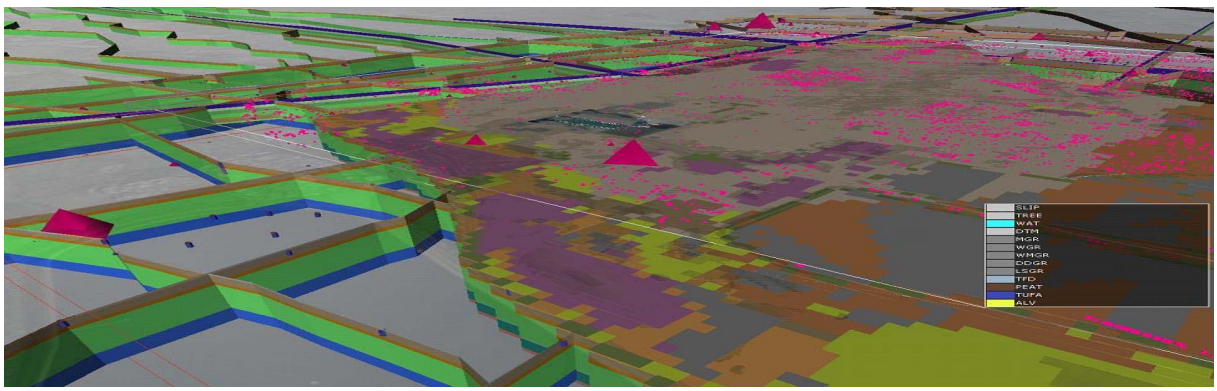

Figure 4: 3D Visualisation in GeoVisionary of property data served through the PropBase "warehouse" system as part of the data architecture.

```
<terms xmlns:xlink="http://www.w3.org/1999/xlink" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
<term xlink:href="http://webservices.bgs.ac.uk/data/services/vocabulary/1.0/vocabularies/DIC_GEOCHRON_RANK/terms/AGE">
<CODE>AGE</CODE>
<STATUS>C</STATUS>
<DATE_ENTERED>18-FEB-04</DATE_ENTERED>
<USER_ENTERED>BGS</USER_ENTERED>
<DATE_UPDATED xsi:nil="true"/>
<DESCRIPTION>PERIOD</DESCRIPTION>
<TRANSLATION>Period</TRANSLATION>
<USER_UPDATED xsi:nil="true"/>
</term>
```

Figure 5. Example XML output from a vocabulary web-service for an entry in a Geochronology dictionary as part of the suite of corporate dictionaries maintained as part of the data architecture

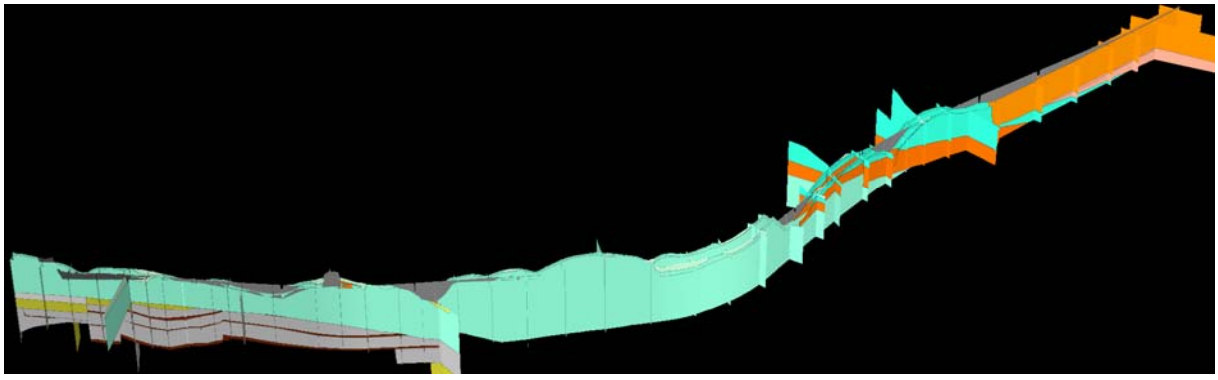

Figure 6: Leeds-York 3D railway model. West to East, x10 vertical Exaggeration. Uses data within the Geoscience Data Hub.

Figure 7: Spatial representation of borehole locations held in BGS's borehole database (>1.3 million). It demonstrates the coverage of the data as the borehole locations closely maps out the map of the UK.

5 Conclusion/Recommendations

Data is a huge asset to organisations especially for GSO's like the BGS and generally for organisations in the geo-environmental science with their data driven research, data products and services delivery requirements. It's critical to have an understanding of your organisation's data and information resources, what they are, where they are stored, managed, standardised and how they can be efficiently accessed by your target audiences. Ensure that they are re-useable and only necessary dupilcation is undertaken whlist preserving the raw and processed states of your data.

Data modelling plays a vital role in achieving all of the above as outlined in the paper the role it has played at the BGS, by modelling our geoscience data it has provided an understanding of our data and assisted us to communicate the information requirements for our geoscientific-environmental research. Data modelling takes different forms and data models provide a vital understanding of the data independent of a subsequent physical implementation.

As a result of the data modelling undertaken at the BGS, allowing for the development of an integrated geoscientific data model, a part of broader Geoscience Data hub, BGS is well positioned for our data driven research, creation and delivery of data products and service. These are objectives very much common to other GSO's and/or geo-environmental organisations and the BGS's data architecture implementation hopefully provides such pointers.

In general, some of benefits to the BGS as a result of the implementation:

- Legal compliance drivers – Freedom of Information, EIR, INSPIRE
- Balanced the benefits of the approach vs. costs in an increasingly harsh funding environment
- Ready to meet the deluge of digital data from data streaming from sensor networks which is the current trend in environmental science to get more data to support our data driven research
- Ensured there are close links between architecture and infrastructure
- Ready for data standardisation-harmonisation, which is becoming critical in science disciplines both in semantics, data models and XML schemas.
- We are in a position to plug in technological solutions/framework into architecture regularly to keep it fit-for-purpose, cost effective, current and evolving.

It's important that we continue to engage in Geosciences, share our data models and implementation types to continue to promote interoperable data systems to enable our data driven research. As part of our Open Geoscience Data models initiative, BGS aims to continue to provide open, ready-to-use database designs that are free for all and to also encourage other organisations to donate models to the resource to meet the needs of the wider environmental community - http://www.bgs.ac.uk/services/dataModels/home.html.

References

1. Connolly, T. and C. Begg (2010). Database Systems: A Practical Approach to Design, Implementation and Management. Fifth Edition. Addison-Wesley.

2. Nayembil, M., A. Richardson, G. Smith, and S. Burden (2014). PropBase Warehouse architecture. In: PropBASE Technical Discussion, Nottingham, (UK) & Copenhagen, (Denmark), 06/03/2014, 14/08/2014 (Unpublished). http://nora.nerc.ac.uk/509988/

3. Nayembil, M. and K. Adlam (2008). Oracle Spatial in British Geological Survey. In: Oracle Spatial Special Interest Group (SIG), British Geological Survey, 07th February 2008. Wimbledon, (UK), UK Oracle User Group (UKOUG). http://nora.nerc.ac.uk/4670/

4. Burke, H. F., L. Hughes, O. J. W. Wakefield, D. C. Entwisle, C. N. Waters, A. Myers, S. Thorpe, R. Terrington, H. Kessler, and C. Horabin (2015). A 3D geological model for B90745 North Trans Pennine Electrification East between Leeds and York. Nottingham, (UK), British Geological Survey, (CR/15/004N) (Unpublished). http://nora.nerc.ac.uk/509777/

5. Watson, C., G. Baker, and M. Nayembil (2014). Open Geoscience Data Models: end of project report. British Geological Survey Open Report (OR/14/061) (Unpublished). http://nora.nerc.ac.uk/508791/